Secrets of RLHF in Large Language Models Part I: PPO

Year: 2023 | *Citations: 225* | *Authors: Rui Zheng, Shihan Dou, Songyang Gao*

Abstract

Large language models (LLMs) have formulated a blueprint for the advancement of artificial general intelligence. Its primary objective is to function as a human-centric (helpful, honest, and harmless) assistant. Alignment with humans assumes paramount significance, and reinforcement learning with human feedback (RLHF) emerges as the pivotal technological paradigm underpinning this pursuit. Current technical routes usually include \textbf{reward models} to measure human preferences, \textbf{Proximal Policy Optimization} (PPO) to optimize policy model outputs, and \textbf{process supervision} to improve step-by-step reasoning capabilities. However, due to the challenges of reward design, environment interaction, and agent training, coupled with huge trial and error cost of large language models, there is a significant barrier for AI researchers to motivate the development of technical alignment and safe landing of LLMs. The stable training of RLHF has still been a puzzle. In the first report, we dissect the framework of RLHF, re-evaluate the inner workings of PPO, and explore how the parts comprising PPO algorithms impact policy agent training. We identify policy constraints being the key factor for the effective implementation of the PPO algorithm. Therefore, we explore the PPO-max, an advanced version of PPO algorithm, to efficiently improve the training stability of the policy model. Based on our main results, we perform a comprehensive analysis of RLHF abilities compared with SFT models and ChatGPT. The absence of open-source implementations has posed significant challenges to the investigation of LLMs alignment. Therefore, we are eager to release technical reports, reward models and PPO codes, aiming to make modest contributions to the advancement of LLMs.